U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

AD-A032 230

A Model for the Analysis of Markovian Decision Processes with Unobservable States and Unobservable Costs

Rand Corp Santa Monica Calif

Oct 72

AD A032230

A MODEL FOR THE ANALYSIS OF MARKOVIAN DECISION PROCESSES WITH UNOBSERVABLE STATES AND UNOBSERVABLE COSTS

James D. Steele*
The Rand Corporation, Santa Monica, California

INTRODUCTION

Consider the Markovian Decision Process (MDP) defined by the following objects;

State Space $S = \{1, 2, 3, \ldots, N\}$, for finite N,

Action Space $A = \{a_1, a_2, \ldots, a_M\}$, for finite M,

Cost Set $C = \{C(i, a_j) : i \varepsilon S, \ a_j \varepsilon A\}$,

Transition Probabilities $= \{q_{ij}(a_K) : i,j \varepsilon S, \ a_K \varepsilon A\}$,

Discount Factor α, such that $o<\alpha<1$.

The problem is to find a policy for taking actions which minimizes the total expected discounted cost over the infinite future, given the initial state of the process.

A stationary policy for (MDP) is defined as a map $f : S \to A$. Howard [2] analyzed (MDP's) having finite state and action spaces and proved that an optimal stationary policy (i.e. a stationary policy which minimizes the total expected discounted cost) always exists. The Howard Policy Improvement Routine is a method by which an optimal stationary policy for (MDP) may be found.

Suppose now, that we are given the (MDP) as defined above, but that we are not allowed to observe the state at any observation point $t = 0, 1, 2, \ldots$. Suppose also, that we are not allowed to observe

---

*The author is an Applied Mathematician at The Rand Corporation.

the cost $C(X_t,a_t)$ at any observation points $t = 0, 1, 2, \ldots$. In other words, the total cost will be assessed at infinity. Finally, suppose that we are allowed to observe the initial probability distribution over the state space S. In this paper we develop a model for analyzing this problem and present some preliminary results. A rather thorough treatment of the problem of unobservable states and unobservable costs for (MDP's) having two states and two actions may be found in Steele [3].

THE MODEL

In an effort to analyze the above problem, we define the following objects.

$$\tilde{S} = \{\text{All probability distributions over } S\}$$

$$= \{\bar{P} = (P_1, P_2, \ldots, P_N) \in E_N : 0 \le P_i \le 1, \sum_{i=1}^{N} P_i = 1,$$

$$i = 1, 2, \ldots, N\},$$

where E_N is N-dimensional Euclidean space and we let P_i be the probability of being in state i;

the set $\tilde{A} = A = \{a_1, a_2, \ldots, a_M\}$; the transition

matrices : $Q(a_K) = [q_{ij}(a_K)]$; and the cost vectors

$\bar{C}(a_K) = (C[1,a_K], \ldots, C[N,a_K])$.

We note that if the distribution over S is $\bar{P}\in\tilde{S}$ and we take action $a_K \in \tilde{A}$, then the new distribution over S will be given by $\bar{P}' = \bar{P}Q(a_K)$. The expected cost, $\tilde{C}(\bar{P},a_K)$, of having the distribution $\bar{P}$ and taking action a_K will be given as the inner product

$$\tilde{C}(\bar{P},a_K) = \langle\bar{P},\bar{C}(a_K)\rangle = \sum_{i=1}^{N} P_i C(i,a_K).$$

At this point, we note that the new distribution $\overline{P}'$ depends only on the current distribution $\overline{P}$ and the current action a_K, i.e. $\overline{P}'=\overline{P}Q(a_K)$, therefore, we see that we now have a new Markovian Decision Process ($\tilde{MDP}$) defined by the objects

State Space $\tilde{S}$ = {all probability distributions over S} ,

Action Space $\tilde{A} = A = \{a_1, a_2, \ldots, a_M\}$, and

Cost Set $\tilde{C} = \{\tilde{C}(\overline{P}, a_K) : \overline{P}\epsilon\tilde{S},\ a_K\epsilon\tilde{A}\}$

Discount Factor α, such that $o<\alpha<1$.

The set of all stationary policies for ($\tilde{MDP}$) is given by $F = \{f : \tilde{S} \to \tilde{A}=A\}$. For any such $f\epsilon F$ and initial state $\overline{P}_o\epsilon\tilde{S}$, the total expected discounted cost is given by

$$V_f(\overline{P}_o) = \sum_{t=o}^{\infty} \alpha^t\, \tilde{C}(\overline{P}_t, f[\overline{P}_t])$$

$$= \sum_{t=o}^{\infty} \alpha^t\, \langle \overline{P}_t, \overline{C}[f(\overline{P}_t)]\rangle, \text{where}$$

$$\overline{P}_t = \overline{P}_o\ Q(f[\overline{P}_o])\ Q(f[\overline{P}_1]) \ldots Q(f[\overline{P}_{t-1}]) \text{ for } t=1,2,3,\ldots .$$

The ($\tilde{MDP}$) as defined above (having uncountable state space and finite action space) belongs to the class of problems analyzed by Blackwell [1]. His analysis showed that an optimal stationary policy always exists and that the Howard Policy Improvement Routine may be extended to this problem. However, in the finite state-finite action problems the set of all stationary policies is finite and, therefore, the Howard Policy Improvement Routine will produce an optimal stationary policy in a finite number of steps. In the uncountable state-finite action problem, the set of all stationary policies is uncountable and, therefore, the Howard Policy Improvement Routine cannot, in general, be used as a method for actually finding an optimal policy.

CONVEX-STATIONARY POLICIES

With F as the set of all stationary policies for $(M\tilde{D}P)$, we define the set $\tilde{F} \subset F$, of all convex-stationary policies for $(M\tilde{D}P)$, as

$$\tilde{F} = \{f \epsilon F : f^{-1}(a_K) \text{ is a convex set for each } a_K \epsilon A\} .$$

CONSTANT SEQUENCE POLICIES

Given the action space $A = \{a_1, a_2, \ldots, a_M\}$, we define $A^N = AxAx\ldots xA$, N-factors, $N = 1,2,3,\ldots$, to be the set of all sequences of length N of elements of A, and we define $A^\infty = AxAxAx\ldots$, to be the set of all infinte sequences of elements of A. For any finite sequence $S \epsilon A^N$, $N \geq 1$, we define the sequence $S^K \epsilon A^{NK}$ to be the sequence $S^K = s, s, \ldots, s$, K-factors, and $S^\infty \epsilon A^\infty$, to be the sequence $S^\infty = s, s, s, \ldots$. For any finite sequences S_1 and S_2 we define $A(S_1;S_2) = A\{S_1, S_2\}$ to be the set consisting of the two action sequences S_1 and S_2 and

$A(S_1;S_2)^\infty = A(S_1;S_2)\, x A(S_1;S_2) x \ldots$, to be the set of all infinite sequences of elements of $A(S_1;S_2)$. For any finite sequence $S = a_1, a_2, \ldots, a_N$, $N \geq 1$, $a_i \epsilon A$, we define

$$L_S(\overline{P}) = \sum_{t=0}^{N-1} \alpha^t \tilde{C}(\overline{P}_t, a_{t+1}),$$

for $\overline{P} \epsilon \tilde{S}$ and $\overline{P}_0 = \overline{P}$, to be the cost of starting at $\overline{P}$ and operating for N time periods when using the $i^{\underline{th}}$ entry in S, $1 \leq i \leq N$, as the action to be taken at the $i^{\underline{th}}$ observation time. We say that $L_S(.)$ is the cost of using the finite sequence S, for all initial $\overline{P} \epsilon \tilde{S}$. If $S \epsilon A^\infty$, we define the constant sequence policy (S), to be the policy which uses the sequence S when starting at any initial $\overline{P} \epsilon \tilde{S}$, we define (A^∞) to be the set of all such policies and we use $V_{(S)}(.)$, in place of $L_{(S)}(.)$, for the cost of the policy (S).

IMMEDIATE RESULTS

LEMMA 1: The cost function, $V_{(r)}$, for any policy $(r) \varepsilon (A^{\infty})$ is linear on $\tilde{S}$.

Proof: Let the sequence $r \varepsilon A^{\infty}$ be given by $r = a_1, a_2, a_3, \ldots$. Now, for any points $\overline{P}$, $\overline{P}'$, $\overline{P}''$ in $\tilde{S}$ such that $\overline{P} = \lambda\overline{P}' + (1-\lambda)\overline{P}''$ for some $\lambda \varepsilon [o,1]$ we have (with $Q(a_o) \equiv$ the Identity operator),

$$V_{(r)}(\overline{P}) = \sum_{t=o}^{\infty} \alpha^t \tilde{C}(\overline{P}_t, a_{t+1})$$

$$= \sum_{t=o}^{\infty} \alpha^t \langle \overline{P}Q(a_o)\, Q(a_1) \ldots Q(a_t), \overline{C}(a_{t+1})\rangle$$

$$= \lambda \sum_{t=o}^{\infty} \alpha^t \langle \overline{P}'Q(a_o) \ldots Q(a_t), \overline{C}(a_{t+1})\rangle$$

$$+(1-\lambda) \sum_{t=o}^{\infty} \alpha^t \langle \overline{P}''\, Q(a_o) \ldots Q(a_t), \overline{C}(a_{t+1})\rangle$$

$$= \lambda V_{(r)}(\overline{P}') + (1-\lambda)\, V_{(r)}(\overline{P}'')$$

or

$$V_{(r)}(\lambda\overline{P}' + [1-\lambda]\overline{P}'') = \lambda V_{(r)}(\overline{P}') + (1-\lambda)\, V_{(r)}(\overline{P}''). \quad \text{Q.E.D.}$$

For any stationary policy f and any $\overline{P}\varepsilon\tilde{S}$, we define the sequence $S(\overline{P},f)\ \varepsilon A^{\infty}$,

$$S(\overline{P},f) = \{a_t\}, \text{ by } a_{t+1} = f(\overline{P}_t),$$

$$\overline{P}_{t+1} = \overline{P}_t Q(f[\overline{P}_t]), \text{ for } t = 0,1,2,\ldots, \text{ and } \overline{P}_o = \overline{P}.$$

We say that $S(\overline{P},f)$ is the sequence generated at $\overline{P}$ when the stationary policy f is used.

LEMMA 2: For any stationary policy f and any $\overline{P}\varepsilon\tilde{S}$, we have

$V_f\,(\overline{P}) = V_{(S[\overline{P},f])}\,(\overline{P})$, where $V_f\,(\overline{P})$ is the cost of using the stationary policy f and starting at $\overline{P}$.

Proof: By definition of $S(\overline{P},f)$. Q.E.D.

THEOREM 1: The optimal cost function is concave on $\mathfrak{S}$.

Proof: Let f^* be optimal. Lemmas 1 and 2 show that at each $\overline{P}\varepsilon\mathfrak{S}$

$$V_{f*}(\overline{P}) = V_{(S[\overline{P},f*])}\,(\overline{P}) = \inf_{S\varepsilon A^\infty} V_{(S)}\,(\overline{P})$$

Therefore, we see that at each point $\overline{P}\varepsilon\mathfrak{S}$, $V_{f*}(\overline{P})$ is the infimum over a set of linear functions and hence V_{f*} is concave. Q.E.D.

Next we prove that the optimal cost function is continuous on $\mathfrak{S}$ by making use of the following representation. Let B be the set of all bounded Baire functions on $\mathfrak{S}$. Define a norm on B by

$$||V|| = \sup_{\overline{P}\varepsilon\mathfrak{S}} |V(\overline{P})|\ , \quad \text{for any } V\varepsilon B.$$

Next, define the operator $U : B\rightarrow B$ by

$$(Uv)\ (\overline{P}) = \min_{a_K\varepsilon A} \{L_{a_K}(\overline{P}) + \alpha V(\overline{P}Q[a_K])\}.$$

In Lemma 3, we state some results presented in Reference [1].

LEMMA 3.

(i) U is a contraction operator

(ii) For any $V\varepsilon B$, the sequence $U_n = U^n V$ converges to the optimal cost function V_{f*}.

(iii) The optimal cost function, V_{f*}, is the unique solution to $UV_{f*} = V_{f*}$.

We now have the following Theorem.

THEOREM 2: The optimal cost function is uniformly continuous on $\mathfrak{S}$.

Proof: For any $u \varepsilon B$, we have

$$u_n = U^n u \to V_{f*} \text{ as } n \to \infty.$$

We also have

$$u_{n+1}(\overline{P}) = \min_{a_K \varepsilon A} \{L_{a_K}(\overline{P}) + \alpha u_n(\overline{P}Q[a_K])\}$$

for $\overline{P} \varepsilon S$. Therefore, we see that since $L_{a_K}(.)$ is continuous for each $a_K \varepsilon A$, each u_n will be continuous if $u = u_o$ is continuous. The convergence of $u_n \to V_{f*}$ is uniform because U is a contraction operator, i.e.

$$||U_n - V_{f*}|| \leq \alpha^n \; ||u - V_{f*}||. \qquad \text{Q.E.D.}$$

BIBLIOGRAPHY

[1] Blackwell, D. (1965). "Discounted Dynamic Programming." Annals of Mathematical Statistics 36, pp. 226-235.

[2] Howard, R.A. (1960). Dynamic Programming and Markov Processes. Cambridge, Mass.; Technology Press, and New York: Wiley.

[3] Steele, J.D., "On Two-state Two-action Markovian Decision Processes With Unobservable States and Unobservable Costs." Unpublished Ph.D. dissertation, Department of Mathematics, University of California at Los Angeles, 1972.